Published online 15 February 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 10* ***3275–3286***
doi:10.1093/nar/gkq073

Optimizing copy number variation analysis using genome-wide short sequence oligonucleotide arrays

Derek A. Oldridge[1], Samprit Banerjee[2], Sunita R. Setlur[3], Andrea Sboner[4] and Francesca Demichelis[1,5,*]

[1]Department of Pathology and Laboratory Medicine, [2]Department of Public Health, Weill Cornell Medical College, 1305 York Avenue, Y 1307 (or box no. 140), NY 10065, [3]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, [4]Molecular Biophysics and Biochemistry Department, Yale University, New Haven, CT 06520 and [5]Institute for Computational Biomedicine, Weill Cornell Medical College, 1305 York Avenue, Y 1307 (or box no. 140), NY 10065, USA

Received October 12, 2009; Revised January 21, 2010; Accepted January 25, 2010

ABSTRACT

The detection of copy number variants (CNV) by array-based platforms provides valuable insight into understanding human diversity. However, suboptimal study design and data processing negatively affect CNV assessment. We quantitatively evaluate their impact when short-sequence oligonucleotide arrays are applied (Affymetrix Genome-Wide Human SNP Array 6.0) by evaluating 42 HapMap samples for CNV detection. Several processing and segmentation strategies are implemented, and results are compared to CNV assessment obtained using an oligonucleotide array CGH platform designed to query CNVs at high resolution (Agilent). We quantitatively demonstrate that different reference models (e.g. single versus pooled sample reference) used to detect CNVs are a major source of inter-platform discrepancy (up to 30%) and that CNVs residing within segmental duplication regions (higher reference copy number) are significantly harder to detect ($P < 0.0001$). After adjusting Affymetrix data to mimic the Agilent experimental design (reference sample effect), we applied several common segmentation approaches and evaluated differential sensitivity and specificity for CNV detection, ranging 39–77% and 86–100% for non-segmental duplication regions, respectively, and 18–55% and 39–77% for segmental duplications. Our results are relevant to any array-based CNV study and provide guidelines to optimize performance based on study-specific objectives.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) and copy number variants (CNVs) are two important components of genomic variation underlying much of human phenotypic diversity. CNVs, recently discovered to exist on a large scale (1,2), are unbalanced rearrangements (e.g. deletions, duplications), operationally defined as segments of DNA larger than 1 kb in size which are present in different numbers of copies between individuals (3–5). CNVs are currently estimated to encompass between 6 and 10% of the human reference genome assembly (Database of Genomic Variants (DGV), http://projects.tcag.ca/variation/) (5,6). The extent to which this common source of variation exists in the human genome is not yet known, due to its incomplete characterization. However, its contribution in the field of evolution (7–9) and disease susceptibility (10) has been demonstrated in the past few years.

When dealing with the detection of CNV loci and with the assessment of individuals' copy number states, one level of complexity arises from the lack of an ideal reference genome; a reference genome is necessary in order to quantify the copy number signal of each individual under study. Another level of complexity when using genome-wide array approaches is defining the appropriate balance between extracting reliable information from the data and maximally exploiting the available information (i.e. specificity versus sensitivity) or applying multiple strategies to separately address false positive and false negative rates. In addition, complex genomic regions (segmental duplications, tandem repeats, architecturally complex CNVs) require special attention to obtain accurate characterization.

*To whom correspondence should be addressed. Tel: +1 646 962 5616; Fax: +1 212 746 8816; Email: frd2004@med.cornell.edu

The most commonly used array based platforms are SNP arrays or hybrid arrays implementing short sequence oligonucleotide probes (e.g. Affymetrix and Illumina use 25- and 50-mers) and long sequence oligonucleotide arrays implementing comparative genome hybridization (CGH) (e.g. Agilent, Nimblegen). The former use short base-pair sequences to capture fragments of DNA and use hybridization intensities to infer copy number without the need of a reference sample cohybridized with the target sample. One advantage of this approach is that it allows for determination of genotypes of SNPs. On the other hand, CGH arrays are known to provide a more accurate determination of DNA copy number due in part to optimization of genomic hybridization (11,12). CGH arrays generally utilize two samples for each experiment in order to empirically compare the amount of DNA present in one sample (target) to a single sample or pooled reference (reference).

The current study identifies problems and suggests possible solutions when dealing with genome-wide array studies involving the characterization of CNVs. Based on a set of 42 individual samples from the HapMap consortium, we present a comparative study of CNV detection by the Affymetrix Genome-Wide Human SNP Array 6.0 platform using previously published data from a custom CNV-targeted CGH platform from Agilent as our standard for comparison. Briefly, the Affymetrix Genome-Wide 6.0 SNP platform comprises ~907 K probes for detecting SNPs in addition to ~946 K non-SNP probes of which 140 K probes specifically target CNV regions. The custom designed Agilent array set comprises ~470 K probes, specifically designed for CNV detection (13). We assess the effects of probe cross-hybridization, of different reference models used to quantify the array signal at time of data processing, and of different data segmentation approaches on CNV assessment for Affymetrix SNP 6.0 data after replicating the Agilent single reference sample setting. For our segmentation analysis, we apply three commonly used data processing algorithms to detect copy number changes (14,15) (http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx). Although the specific performance numbers that we report reflect the present study design, we focus on observations and results that are conceptually applicable to any CNV oriented study.

MATERIALS AND METHODS

Sample selection

All samples considered in this study come from the International HapMap Project (16). For marker QC analysis, data was considered from 270 HapMap individuals, consisting of 30 trios from Utah residents of European ancestry (CEU), 30 trios from the Yoruba in Ibadan, Nigeria (YRI), 45 unrelated Japanese in Tokyo (JPT), and 45 unrelated Han Chinese in Beijing (CHB). For cross platform comparison, a 30 sample subset of these 270 individuals was selected (see Supplementary Table S1), including 10 unrelated CEU, 10 unrelated YRI, 5 unrelated JPT and 5 unrelated CHB individuals. This sample set reflects the study from Perry *et al.* (13). An additional 12 HapMap samples profiled using the Agilent Technologies Human Genome CNV microarray set (G4423B, AMADIDs 018897 and 018898) were considered.

Comparative genomic hybridization data

Data for the 30 HapMap samples analyzed with the custom Agilent array set considered in Perry *et al.* were obtained from the Gene Expression Omnibus under accession number GSE9831. Sample NA10851 was used as the reference sample. Data preprocessing is described in ref. (13). An additional 12 HapMap samples were analyzed using a similar data analysis workflow; sample NA15510 was used as a reference. See Supplementary Methods section.

Short sequence oligo-nucleotide data

Raw HapMap data generated using the Affymetrix Genome-Wide Human SNP Array 6.0 were obtained from Affymetrix (Affymetrix, Santa Clara, CA). Raw data was preprocessed according to the Affymetrix CN5 method included in Affymetrix Power Tools (APT) (http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx). Similar to Perry *et al.* (13), Affymetrix marker sequence alignment against the human genome reference sequence (hg18) was performed to identify all locations with perfect (25-mer) and imperfect (23–24-mer) matches. Only markers which map to a single perfect match and fewer than four 24-mer match locations were retained for CNV detection analysis. Briefly, a total of 15 631 such SNP markers (out of

Table 1. Summary statistics for positive copy number deviations (gains) called by segmentation of Affymetrix data across the entire genome

Algorithm	Parameters	Gain count mean	Gain size (bp)			Gain marker count			Gain $\log_2$ intensity ratio		
			Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
CBS	$\alpha = 0.002$	69	59 543	7615	176 276	38	18	112	1.06	0.73	0.72
CBS	(Default) $\alpha = 0.010$	82	53 922	6815	165 482	34	15	98	1.03	0.74	0.69
CBS	$\alpha = 0.010$, sdundo = 2	52	22 266	4082	131 444	17	6	133	1.35	1.25	0.68
CBS	$\alpha = 0.050$	127	40 620	4665	140 259	24	8	73	0.99	0.80	0.62
GLAD	(Default) $d = 6$	91	34 713	2785	141 832	24	4	104	1.43	1.31	0.83
GLAD	$d = 12$	53	59 608	5152	189 243	39	8	136	1.50	1.49	0.88
HMM	(Default)	116	51 617	8046	361 000	26	11	77	0.97	0.75	0.60

905 215) and 26 052 such CN markers (out of 908 226) were excluded according to these criteria. Markers residing within segmental duplications were more likely to fail the above criteria than those which do not reside within segmental duplications (OR = 15.6, $P < 2 \times 10^{-16}$). Importantly, of the 41 730 markers annotated as residing within segmental duplication regions (UCSC annotation as of November 2009), only 9366 (22.4%) are flagged for exclusion based on the above marker alignment analysis, indicating that many segmental duplication markers remain in order to detect copy number variation at segmental duplication loci (see Supplementary Data in Supplementary Methods section). For more detail on platform design and sequence alignment and marker response evaluation, see Supplementary Data in Supplementary Methods section.

Reference models

Affymetrix $\log_2$ intensity ratio data was considered with respect to one of two possible references: (i) a 210 parental HapMap reference model (built on a marker basis by taking the median values across the 210 samples) or (ii) a single sample reference based on either NA10851 or NA15510, to reflect the original Agilent CGH array data. Transformation of $\log_2$ ratio data from the original multiple sample reference model to a single sample reference takes advantage of the following logarithmic identity: $\log_2(\text{Target}/210\text{HapMap}) - \log_2(\text{SingleSample}/210\text{HapMap}) = \log_2(\text{Target}/\text{SingleSample})$.

Data segmentation

Three commonly used segmentation algorithms were applied to the Affymetrix data. Gain and Loss Analysis of DNA (GLAD, R package from Bioconductor) (14) was applied with d equal to 6 (default) and to 12 (d controls the penalty of adding markers to a segment). Circular Binary Segmentation (CBS, 'DNAcopy' package from Bioconductor) (15) was applied with $\alpha = 0.002$, 0.010 (default), and 0.050. A Hidden Markov Model (HMM) approach was applied using the HMM algorithm provided in the standard APT workflow with default parameters.

RESULTS

Segmentation results

We evaluated the summary statistics for the count, size in terms of base pairs and number of markers, and amplitude of gain and deletion deviations produced by segmentation of Affymetrix across 42 HapMap samples (see Tables 1 and 2). Upper and lower bounds of ± 0.15 were applied for distinguishing real copy number gain and loss from no change, respectively. Notably, CBS with $\alpha = 0.050$ and HMM tend to generate the largest number of gain segments, with an average of 127 and 116 gains per individual, respectively, whereas CBS with $\alpha = 0.010$ and sdundo = 2, and GLAD with $d = 12$ generate the fewest, with 52 and 53 average gains, respectively. A similar trend is observed for deletions, where CBS with $\alpha = 0.050$ and HMM generate 160 and 168 deletions per individual, respectively, whereas CBS with $\alpha = 0.010$ and sdundo = 2, and GLAD with $d = 12$ generate the fewest, with 75 and 68 deletions, respectively. Across algorithms, median gain sizes range from 2.8 to 8 kb and from 4 to 18 markers, whereas median deletion sizes range from 1.3 to 4.4 kb and from 3 to 6 markers. Median $\log_2$ intensity ratio amplitude varies from 0.73 to 1.49 for gains and from −1.00 to −1.65 for deletions. Average segment size and amplitude are generally inversely correlated with the number of segments that are called.

Performance measures: false negative and false positive rate

False negative rates (FNR) and false positive rates (FPR) for segmented Affymetrix SNP 6.0 data are defined in reference to the sample level CNVs determined by Agilent aCGH, which we consider our 'gold standard' for comparison. We defined a false negative as a CNV reported in Agilent data that is not detected as a copy number deviation by segmentation of Affymetrix data, and FNR is the proportion Agilent CNVs which are not concordantly detected on the Affymetrix platform, evaluated on a sample basis. Analogously, a false positive is a copy number deviation determined from segmentation of Affymetrix data that does not correspond to an Agilent CNV, and FPR is the proportion of Affymetrix copy number deviations which are not concordantly reported in the Agilent data, evaluated on a sample basis.

Only Affymetrix segments of 10 or more markers and an absolute mean $\log_2$ intensity ratio of 0.15 or more are considered to be significant copy number deviations for evaluating FNR and FPR. To account for differential genomic coverage of the Affymetrix and Agilent platforms, only Agilent CNVs with at least 10 markers of Affymetrix overlap are considered when calculating

Table 2. Summary statistics for negative copy number deviations (deletions) called by segmentation of Affymetrix data across the entire genome

Algorithm	Parameters	Deletion count mean	Deletion size (bp)			Deletion marker count			Deletion $\log_2$ intensity ratio		
			Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
CBS	$\alpha = 0.002$	83	58 029	4280	214 436	22	7	47	−1.27	−1.28	0.69
CBS	(Default) $\alpha = 0.010$	100	49 885	3787	191 598	20	6	42	−1.21	−1.19	0.67
CBS	$\alpha = 0.010$, sdundo = 2	75	15 748	2085	73 529	10	3	27	−1.45	−1.44	0.59
CBS	$\alpha = 0.050$	160	36 158	2947	166 297	14	4	34	−1.13	−1.09	0.60
GLAD	(Default) $d = 6$	119	29 512	1348	153 514	13	3	39	−1.45	−1.42	0.69
GLAD	$d = 12$	68	52 372	1711	224 355	21	3	55	−1.62	−1.65	0.74
HMM	(Default)	168	28 379	4360	112 994	13	6	29	−1.23	−1.00	0.71

FNR, and only Affymetrix copy number deviations with at least five markers of Agilent coverage are considered when calculating FPR. Supplementary Table S2 reports the statistics of the number of segments and CNVs included in each analysis. When determining sample level concordance between the Affymetrix deviation profile and the Agilent CNV profile, we apply an overlap criterion for a CNV to be considered concordantly detected by both the Affymetrix and Agilent platforms. When calculating FNR (FPR), at least 50% of the overlapping Affymetrix (Agilent) probes must be called as Affymetrix copy number deviations (Agilent CNVs), with concordant direction. This overlap criteria based on the percentage of probes as opposed to size percentage takes uneven marker spacing into consideration between the two platforms.

FNR and FPR are evaluated on an algorithmic basis in order to assess performance in terms of sensitivity and specificity for detecting CNV. In addition to considering the overall performances, we considered various subcategories of CNVs defined by the number of markers or by the average $\log_2$ ratios in order to assess the effects of CNV size, direction and amplitude on performance. Specifically, for FNR, we subdivided by CNVs with 2–4, 5–10, or $\geq$11 Agilent probes; by deletion CNVs with mean $\log_2$ ratio ≤ -0.75, or > -0.75 and ≤ -0.15; and by gain CNVs with mean $\log_2$ ratio ≥ 0.15 and < 0.75, or ≥ 0.75. For FPR, we subdivided by CNVs with 10–19, 20–29, or $\geq$30 Affymetrix probes; and by deletion and gain CNVs with mean $\log_2$ ratios as for FNR.

Correction methods to account for reference model/sample

Before presenting the FNR and FPR values for the direct comparison of Affymetrix versus Agilent data by simulating the same experimental design, we quantitatively examine the effect of the reference model/sample on FNR and FPR. A dual channel Agilent aCGH experiment uses a reference based on a single sample or pooled samples which are co-hybridized with the target sample during array hybridization, whereas in a single channel Affymetrix experiment, only the target sample is hybridized to a given array and a reference model is computationally applied based on median marker signal across multiple samples following hybridization. This difference is a source of systematic discordance and reflects a general problem introduced in inter-study CNV data comparison (reference sample/model effect). For example, in a single reference sample design, a deletion in the single sample reference would result in the appearance of a copy number gain in any target samples with no copy number change at that locus, and it would appear as if there was no copy number change for any target samples which show the same deletion (Figure 1A). On the other hand, for low to moderate CNV frequency, a multiple sample median reference model would correctly reflect the no copy number change state as the reference state and allow for correct evaluation of the target sample state (Figure 1B and C).

We considered three methods to correct for systematic copy number profile discordance due to reference differences; the first approach is a stringent and simple utilization of CNV frequency information (does not require the availability of the raw data), the second one utilizes the Agilent single channel data (raw data needed), and the third one transforms the Affymetrix data to mimic the Agilent experimental design (requires availability of single reference sample Affymetrix data). The results are summarized in Figure 2.

Frequency-based approach. In general, rare CNVs present in the single sample reference will lead to an unusually high number of target samples showing the 'opposite' CNV state. We set a frequency threshold of $\geq$66% and then ran the comparison on the remaining 93.6% of CNV loci (2759) from Perry *et al*. (13). On a sample level, this stringent criterion resulted in 56% of CNV exclusion for FNR evaluation and in 28–74% of CNV exclusion for FPR. The frequency based approach is very simple and does not require specific data processing; however, it dramatically reduces the number of loci for inter-study comparison.

Agilent single channel approach. Single channel signal intensities were analyzed to select all Agilent detected CNVs where the reference sample is estimated to have two copies. For each CNV and each experiment, we compute the median reference channel signal intensity across all probes that fall in this CNV. The histogram of these median intensities is shown in Supplementary Figure S3. The first three peaks on the signal intensity distribution correspond to zero, one and two copies, and we conservatively estimate that the reference sample has two copies in CNVs with median signal intensities ranging between 450 and 700 counts. Only these CNVs were used for comparison with Affymetrix data.

Single sample reference model for Affymetrix data. Affymetrix data are computationally transformed to mimic the Agilent single reference sample design (either NA10851 or NA15510 based on the original Agilent data, see 'Materials and Methods' section).

This correction method does not exclude any CNV regions in the FNR or FPR analysis. The transformed Affymetrix data are re-segmented, and FNR and FPR are considered as specified above based on the new Affymetrix copy number deviation profile determined by segmentation.

Figure 2 shows FNR and FPR performance of these three correction methods with respect to the uncorrected FNR and FPR. Notably, both the *frequency-based approach* and the *agilent single channel approach* led to improved FNR (44 and 52%, respectively, relative to 65% uncorrected) and FPR (38 and 33%, respectively, relative to 42% uncorrected). However, the best performance is observed by simply transforming the Affymetrix data so that its implicit reference is in agreement with the Agilent data (37% FNR and 20% FPR, nearly half of the uncorrected rates).

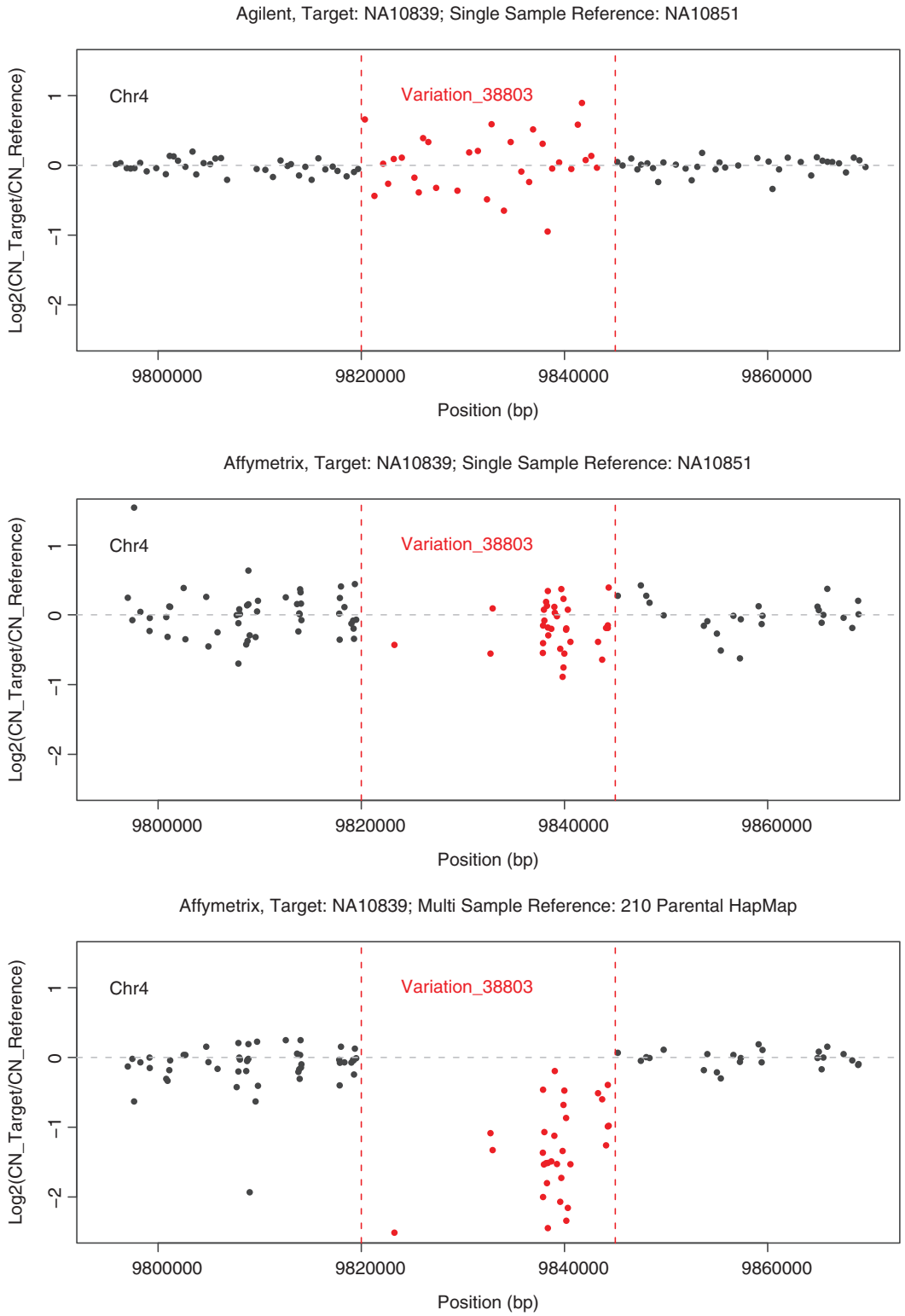

Figure 1. Impact of the reference model copy number data for sample NA10839 at polymorphic locus chr4:9 823 254–9 844 366 (Variation_38803 from DGV, in red) and flanking areas (in black), as obtained by three different approaches. *Top*: Dual channel Agilent data. By design, sample NA10851 was used as reference (single sample reference). NA10839 appears to have no change in copy number. This is a spurious effect, due to the presence of homozygous deletions at Variation_38803 in both NA10839 and NA10851. An increased signal-to-noise ratio is appreciable within the polymorphic locus, where the root mean square of the $\log_2$ ratio is 3.97 times higher than the flanking areas. *Middle*: Single channel Affymetrix data. Hybridization intensity data of sample NA10839 is analyzed with respect to NA10851 data (single sample reference). Like the top panel, NA10839 appears to have no change in copy number at Variation_33803. The root mean square of the $\log_2$ ratio within the polymorphic locus is 1.25 times higher with respect to the flanking areas. *Bottom*: Single channel Affymetrix data. Hybridization intensity data of sample NA10839 is analyzed with respect to a multi-sample reference model based on the median signal across 210 parental HapMap. The $\log_2$ intensity data shows the deletion at Variation_33803. The median value across the markers within the polymorphic locus is −1.36.

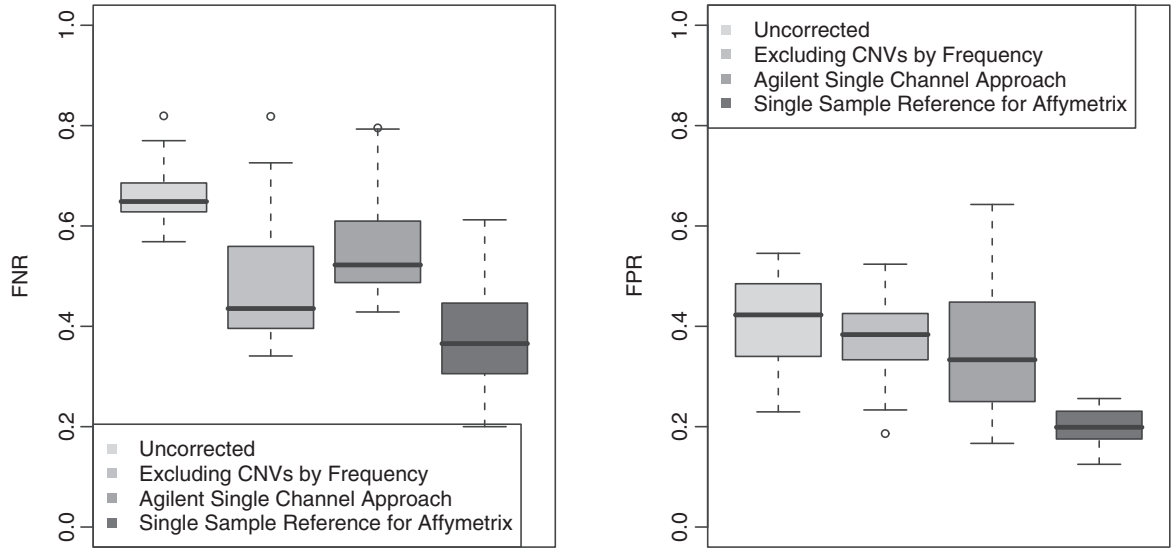

Figure 2. Correction methods for evaluating FNR and FPR. FNR and FPR evaluated by comparing Affymetrix copy number deviation profiles with an Agilent gold-standard across 30 HapMap samples prior to correcting for different reference models implicit in Agilent and Affymetrix $\log_2$ intensity ratio data and after applying three different correction methods. For uncorrected FNR and FPR, Agilent data is considered with respect to a single HapMap sample (NA10851) reference, whereas Affymetrix data is considered with respect to a multiple sample, 210 parental HapMap median reference model. The first correction method excludes regions from consideration where 20 or more of the 30 samples have a CNV as detected by Agilent in a given region, since high frequency CNVs identified in the Agilent data are likely to arise from rare CNVs which are present in the single sample reference. The second correction method is based on analyzing Agilent single channel (reference channel) data directly to identify regions which are likely to be altered in the single reference sample. The final correction method transforms Affymetrix $\log_2$ intensity ratio values so that they are implicitly computed with respect to a single sample, NA10851 reference, effectively bringing the Affymetrix and Agilent reference models into agreement. Both the CNV frequency and reference channel correction methods improve FNR and FPR performance, where the CNV frequency method is better for FNR and the reference channel method is better for FPR. However, simply transforming the Affymetrix reference to be in agreement with the single sample reference implicit in the Agilent data yields the lowest FNR and FPR, and therefore the highest concordance between the Affymetrix and Agilent platforms. All segmented Affymetrix data are obtained from the CBS algorithm with $\alpha = 0.010$, though the same qualitative result is observed for the other segmentation algorithms which we consider.

It is important to note that single sample reference based analyses of Affymetrix data is commonly performed in the context of paired tumor-versus-normal analysis, but is rarely applied in the context of germline CNV assessment. However, in the context of our study design, this last approach directly removes the confounding effect of different reference models used for our Affymetrix versus Agilent comparison and allows us to more accurately assess segmentation algorithm performance without the need to exclude CNV regions, which would limit the scope of our FNR and FPR analysis. Therefore, all further FNR and FPR analysis is performed with the Affymetrix $\log_2$ intensity ratio data transformed to the appropriate single sample reference.

Segmental duplication regions

Segmental duplication regions are known to be enriched for deletion and duplication events, especially mediated by non-allelic homologous recombination (NAHR) and are therefore of interest in CNV studies. However, segmental duplication regions are more difficult to handle and need dedicated attention (6,17). In fact, it becomes more difficult to distinguish a one copy change in the target data, since the expected $\log_2$ intensity ratio decreases as a function of increasing reference copy number. We illustrate this theoretical decrease in signal schematically in Figure 3A [e.g. a one copy gain at a segmental duplication is harder to detect since $\log_2(5/4) < \log_2(3/2)$]. We empirically evaluated the distribution of marker level $\log_2$ intensity ratios across 270 HapMap samples at known CNVs considered in McCarroll *et al.* (4) (Figure 3B–E), distinguishing between segmental duplication and non-segmental duplication regions. Annotation of segmental duplications was obtained from the UCSC Genome Browser. As expected, the signal distribution differentiation among different copy number states is reduced in segmental duplication regions, especially for copy number losses (P-value $< 2 \times 10^{-16}$, for CN = 0, 1 and 3; P-value = 0.0027, for CN = 4). This suggests that the assessment of CNV states within segmental duplication regions requires specific thresholds and needs to be treated separately from non-segmental duplication regions.

Supplementary Table S3 includes the summary statistics of non-segmental duplication CNVs included in the analysis per segmentation approach. Supplementary Table S4 summarizes the average per sample counts of CNVs detected by the Affymetrix and Agilent platforms, per every CNV subcategory considered in this study.

Comparison of CNV detection approaches

In this section we separately report on the analysis of non-segmental duplication and segmental duplication regions and discuss the main differences. Detailed comparison of CNV detection approaches is summarized in Tables 3 and 4. For non-segmental duplication regions, CBS with $\alpha = 0.050$ and HMM exhibit the greatest overall sensitivity (23% FNR), whereas GLAD and CBS

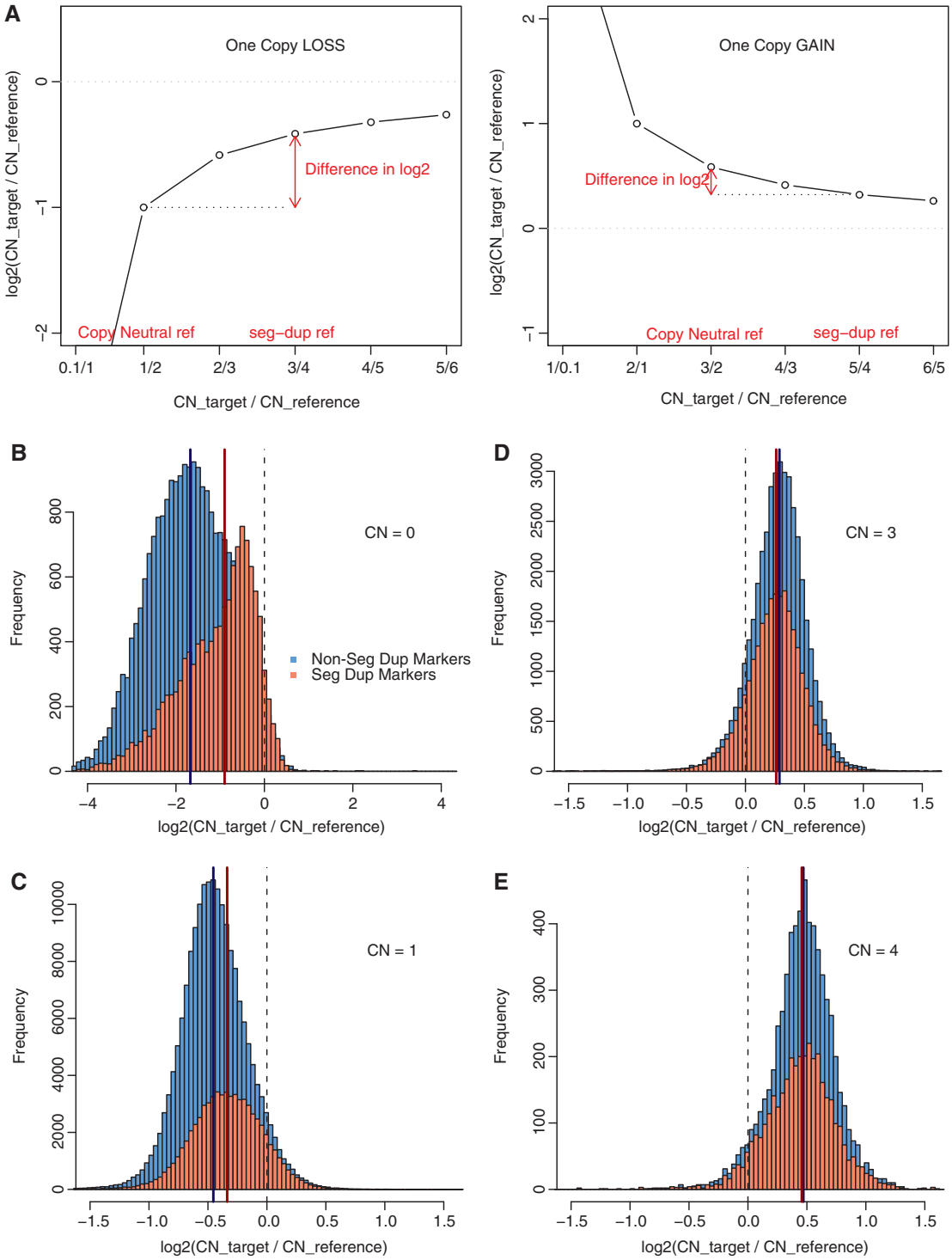

Figure 3. Schematic and empirical data of $\log_2$ intensity ratios in the context of segmental duplication and non-segmental duplication regions. (**A**) Schematic of $\log_2$ intensity ratios for one copy number change CNVs with respect to different reference states. Red vertical arrows indicate the differences between copy neutral reference and segmental duplicate (seg-dup) reference for one copy loss (left panel) and one copy gain (right panel). (**B–E**) Distribution of $\log_2$ intensity ratios for Affymetrix SNP 6.0 Array markers within knwn polymorphic CNV loci, distinguishing between segmental duplication annotated regions (colored red) and non-segmental duplication regions (colored blue). Data is considered from the subset of 270 HapMap samples showing copy number variation with respect to a copy neutral (CN = 2) median reference at each CNV locus. The cases of homozygous deletion (CN = 0, panel B), hemizygous deletion (CN = 1, panel C), one copy gain (CN = 3, panel D), or two copy gain (CN = 4, panel E) are assessed independently based on sample level CN state calls reported in McCarroll *et al*. Vertical red and blue lines identify the medians of the distributions. In the case of homozygous deletion (CN = 0), markers which reside within segmental duplication regions exhibit a median $\log_2$ intensity ratio of −0.90, whereas those markers which do not reside within segmental duplication regions exhibit a median $\log_2$ intensity ratio of −1.68 ($P < 2 \times 10^{-16}$). For the case of hemizygous deletions, the median $\log_2$ intensity ratio values are −0.34 versus −0.45, respectively ($P < 2 \times 10^{-16}$); for one copy gain, 0.26 versus 0.29, respectively ($P < 2 \times 10^{-16}$); and for two copy gain, 0.46 versus 0.47, respectively ($P = 0.0027$). Overall, markers residing within segmental duplication regions exhibit less sensitivity to detect copy number variation, as indicated by a smaller absolute median $\log_2$ intensity ratio for each case considered.

Table 3. Evaluation of difference between FNR calculated separately for segmental duplication and non-segmental duplication CNVs

Algorithm	Parameters	FNR for segmental duplication CNVs			FNR for non-segmental duplication CNVs			*P*-value
		Median	Lower 95%	Upper 95%	Median	Lower 95%	Upper 95%	
CBS	$\alpha = 0.002$	0.53	0.34	0.79	0.28	0.06	0.59	$1.02E-09$
CBS	(Default) $\alpha = 0.010$	0.50	0.32	0.77	0.26	0.08	0.55	$2.46E-10$
CBS	$\alpha = 0.010$, sdundo = 2	0.82	0.64	0.92	0.61	0.29	0.85	$7.12E-09$
CBS	$\alpha = 0.050$	0.45	0.27	0.74	0.23	0.10	0.48	$4.77E-10$
GLAD	(Default) $d = 6$	0.68	0.48	0.86	0.55	0.31	0.80	0.000201
GLAD	$d = 12$	0.70	0.46	0.86	0.61	0.34	0.82	0.004025
HMM	(Default)	0.50	0.36	0.76	0.23	0.06	0.44	$5.74E-13$

Table 4. Evaluation of difference between FPR calculated separately for segmental duplication and non-segmental duplication CNVs

Algorithm	Parameters	FPR for segmental duplication CNVs			FPR for non-segmental duplication CNVs			*P*-value
		Median	Lower 95%	Upper 95%	Median	Lower 95%	Upper 95%	
CBS	$\alpha = 0.002$	0.29	0.05	0.50	0.06	0.00	0.13	$6.17E-12$
CBS	(Default) $\alpha = 0.010$	0.32	0.05	0.49	0.10	0.00	0.20	$7.45E-12$
CBS	$\alpha = 0.010$, sdundo = 2	0.15	0.00	0.43	0.03	0.00	0.15	$3.95E-06$
CBS	$\alpha = 0.050$	0.36	0.10	0.49	0.14	0.05	0.29	$3.28E-10$
GLAD	(Default) $d = 6$	0.24	0.03	0.52	0.03	0.00	0.09	$1.69E-13$
GLAD	$d = 12$	0.23	0.04	0.50	0.00	0.00	0.10	$8.37E-14$
HMM	(Default)	0.33	0.05	0.64	0.11	0.02	0.25	$2.35E-09$

with sdundo = 2 exhibit the lowest (61% FNR) (Table 3). Across subcategories, the best performances are seen for high amplitude gains or deletions (absolute $\log_2$ intensity ratio > 0.75) where FNR ranges ~10–50% across algorithms. Significantly poorer performance is seen in the 2–4 Agilent marker, small CNV category (50–90% FNR) and in the low amplitude gain (40–90% FNR) and low amplitude loss categories (75–100% FNR). As expected, there is a trend toward lower FNR with increasing CNV size and amplitude. There is also a trend toward lower FNR as the CBS parameter, α, increases, or as the GLAD parameter, d, decreases, consistent with their expected effects on sensitivity. Though CBS is generally more sensitive than GLAD, it is noteworthy that running CBS with parameter sdundo = 2 yields similar performance to GLAD. A summary of FNR performance evaluated across all algorithms, parameters, subcategories, and samples considered, for non-segmental duplication CNVs, is compiled in Figure 4A.

The trend for overall FPR is essentially the reverse of overall FNR, where CBS with $\alpha = 0.050$ and HMM show the lowest specificity (~12% FPR), and GLAD or CBS with sdundo = 2 show the highest (nearly 0% FPR) (Table 4). The most specific algorithms perform with 100% or near 100% specificity across all CNV categories. For algorithms with non-zero FPR, specificity typically ranges 80–90% across subcategories. One important comment is that FPR is lower than FNR across nearly all CNV subcategories regardless of algorithm, indicating that the sensitivity associated with these algorithms does not incur a large penalty to specificity. However, CBS with $\alpha = 0.050$ and HMM do show nearly twice the FPR associated with CBS with $\alpha = 0.002$. When FPR is greater than zero, there is an apparent trend towards lower FPR with increasing CNV size and amplitude. Similar to the results for FNR, running CBS with sdundo = 2 yields comparable FPR performance to GLAD. A summary of FPR performance evaluated across all algorithms, parameters, subcategories, and samples considered, for non-segmental duplication CNVs, is compiled in Figure 4B.

Lower performance is observed for segmental duplication CNVs, where FNR increases 9–27% and FPR increases 12–24% relative to the results considered for non-segmental duplication CNVs (Tables 3 and 4). However, the relative ranking of segmentation algorithms based on sensitivity and specificity is otherwise unchanged and similar trends based upon CNV size and amplitude are observed. One notable exception for FNR is in the 11 or more Agilent marker, large CNV category, where higher FNR is observed relative to the 5–10 Agilent marker, medium-sized CNV category. This observation is consistent with the presence of large segmental duplication CNVs which are detected at low-amplitude by the Agilent platform due to a higher associated reference copy number. Comparing segmental duplication and non-segmental duplication CNVs, FNR increases in the 11 or more Agilent marker category by ~20–30% across algorithms, and it is this category that drives the overall increase in FNR. There is a statistically significant difference (Wilcoxon test, $P < 0.05$) between overall FNR evaluated for segmental duplication CNVs versus non-segmental duplication CNVs across all algorithms. FPR comparisons across algorithms are also significantly different (Wilcoxon test, $P < 0.05$). The graphical representation of FNR and FPR performance evaluated across all algorithms, parameters, subcategories and

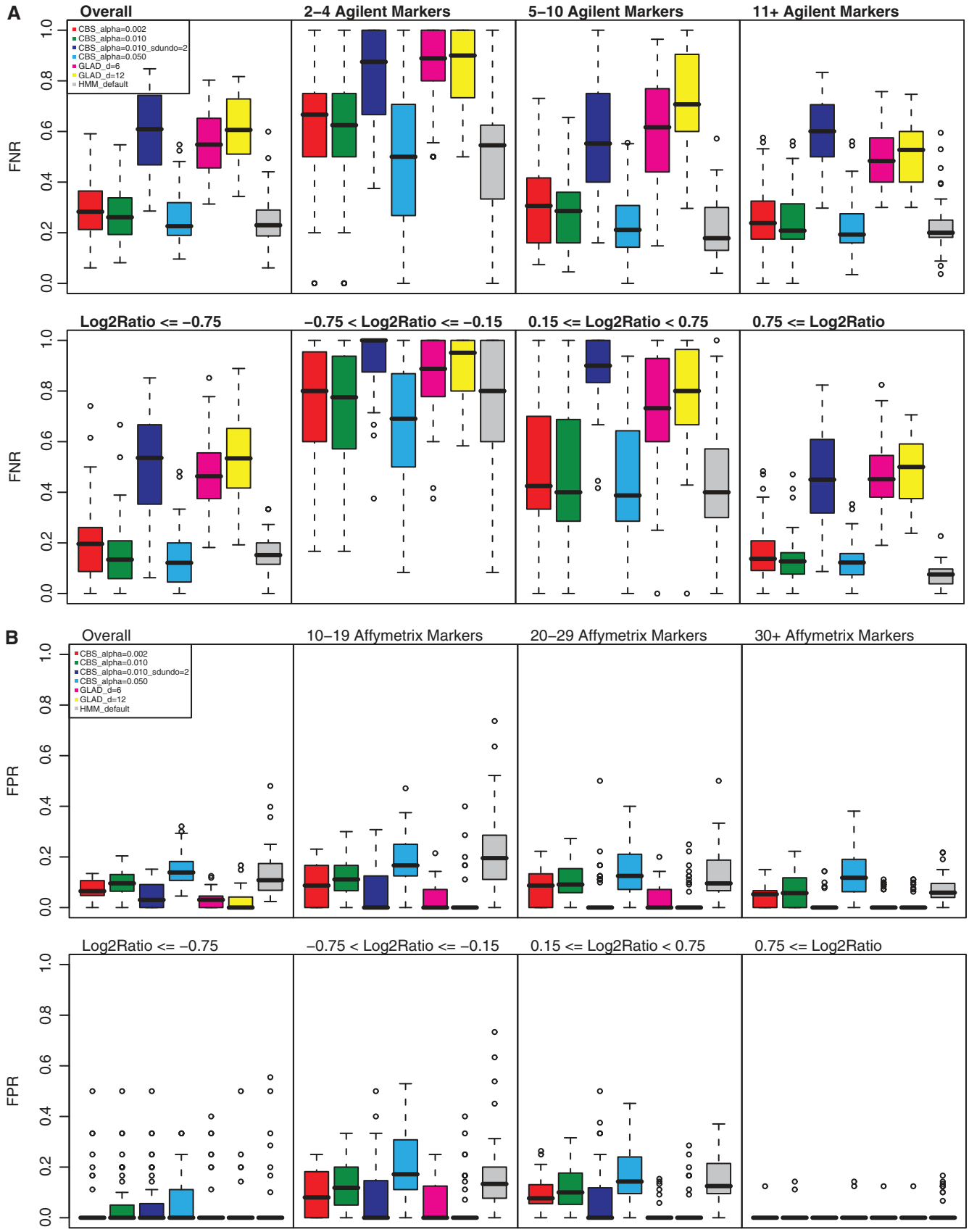

Figure 4. (**A**) FNR for non-segmental duplication CNVs. FNR evaluated by comparing Affymetrix copy number deviation profiles with an Agilent gold-standard across 42 HapMap samples, considering only non-segmental duplication CNVs. Overall and across nearly all subcategories, CBS with $\alpha = 0.050$ and HMM exhibit the greatest sensitivity, whereas CBS with $\alpha = 0.010$ and sdundo = 2 and GLAD with $d = 12$ exhibit the least.

samples considered, for segmental duplication CNVs alone, is compiled in Supplementary Figures S4 and S5.

In summary, we find that different combinations of segmentation algorithms and their associated parameters yield a range of specificities and sensitivities for detecting copy number variation. Although tuning algorithm-specific parameters has slight effects on sensitivity and specificity, CBS and HMM are inherently more sensitive for detecting copy number variation whereas GLAD is inherently more specific. However, merging segments based on a standard deviation threshold after CBS segmentation yields results which are comparable to GLAD, indicating that post-segmentation processing can be used when greater specificity is desired. Although no algorithm and parameter set is able to detect even a simple majority of small CNVs with relatively poor Affymetrix marker coverage or low amplitude CNVs, CBS and HMM will perform comparatively better than GLAD in detecting such low-signal CNVs. On the other hand, we find that the GLAD algorithm is inherently more conservative when calling copy number segments, with near zero FPR across all CNV categories considered in this study. GLAD may therefore be best suited for applications where specificity is at a premium. The appropriate segmentation algorithm will largely depend on the particular needs of the application at hand.

DISCUSSION

Ongoing studies aimed at CNV characterization, assessment of disease/phenotype association and evolutionary investigation rely on array-based approaches. DGV collects and organizes the data generated in the research community related to CNVs and insertions and deletions. It is widely used in the genomic community both to retrieve information on a specific region or CNV from multiple studies and to retrieve genome-wide information from one or more studies for comparison purposes. Inter-study comparisons are delicate due to multiple factors related to different platforms and experimental designs. In particular, differences in platform spatial resolution (inter-marker distance), dynamic range, reference model, platform strengths and weaknesses must all be taken into account together with the study data processing approaches. Furthermore, complex genomic regions such as segmental duplications, tandem repeats, and complex CNV areas require special attention. In addition, the size and the ethnic background of each study cohort need to be considered especially when assessing and comparing CNV frequencies of rare CNVs and low-frequency polymorphisms (18). A recent paper from Shaikh *et al.* (5) showed that the reliable assessment of low frequency CNVs in healthy individuals (controls) is particularly crucial when dealing with genotype–phenotype association studies.

In the current study, we quantitatively evaluate the impact of different experimental and data processing alternatives and choices in the context of CNV detection and copy number assessment by defining strict performance measures when comparing data generated with a genome-wide short sequence oligonucleotide array (Affymetrix) with a long sequence oligonucleotide CGH array custom designed for CNV detection (Agilent). First, we evaluate the impact of using different reference models (single- versus multi-sample reference model) when detecting and assessing CNV states; we measure that implementing different reference models in the analysis pipeline can lead to overall FNR and FPR differences as high as 25 and 20%. We suggest that overall a multi-sample reference model is a preferable solution in the context of CNV discovery, as it tends to avoid misleading results, such as the absence of a CNV call when the target and the reference sample both possess the same CNV state different than the normal state (see Figure 1). The main caveat in using the multi-sample reference model has to do with assessing the copy number state at polymorphic loci with moderate to high incidence (around and above 50%). For instance, a homozygous/hemizygous deletion in 51% of the population will result in a multi-sample reference model reflecting the deletion, therefore resulting in no detectable change for each target sample with the deletion and in a gain for each target sample with two copies. For known CNV loci, computational approaches, like 'Canary' (19), can circumvent the reference model problem by taking advantage of the relative differences between adjacent sample signal clusters, which are independent of the applied reference model. Due to reference model differences between studies, publicly accessible CNV information can be misleading; one example is shown in Supplementary Figure S6 where a region known to undergo deletions in 40–50% of Caucasian population is annotated in DGV as harboring gains in one-third of the available studies, including the paper from Perry *et al.* (13). It is interesting to note that many studies aimed at characterizing the somatic

Figure 4. Continued
As expected, there is a general trend toward lower FNR with increasing CNV size or amplitude, since such CNVs will benefit from an increased signal-to-noise ratio and will be easier to detect. Comparing CBS and GLAD, the effect of varying the CBS parameter α or the GLAD parameter have modest effects on FNR in relation to changing algorithms. However, applying the CBS parameter sdundo = 2 yields performance which is comparable with GLAD. (**B**) FPR for non-segmental duplication CNVs. FPR evaluated by comparing Affymetrix copy number deviation profiles with an Agilent gold-standard across 42 HapMap samples, considering only non-segmental duplication CNVs. Overall and across nearly all subcategories, CBS with $\alpha = 0.010$ and sdundo = 2, and GLAD with $d = 12$ exhibit the greatest specificity, whereas and CBS with $\alpha = 0.050$ and HMM exhibit the least, inverting the trend for sensitivity observed in the FNR analysis. There is a general trend toward lower FPR with increasing Affymetrix copy number deviation size or amplitude. This is expected, because nearly all real CNVs which can be concordantly detected by the two platforms should be detected by the Agilent platform. Therefore, false positives primarily represent random noise of the Affymetrix platform, and it is unlikely that progressively larger or higher amplitude copy number deviations will be called by segmentation from random noise. Comparing CBS and GLAD, the decreased sensitivity of GLAD observed in the FNR analysis is offset somewhat by near perfect specificity observed for the GLAD algorithm.

aberration of tumor cells implemented array based approaches in the past without experiencing reference model issues, since germline DNA from the same individual provides the ideal reference for somatic lesion detection (20,21).

To ensure the most comparable settings between the Affymetrix and the Agilent data, we used a single sample reference model reflecting the original data set-up to evaluate the performance of various data analysis approaches. We first show that segmental duplication regions are significantly harder to process for CNV detection. Then we demonstrate that different analysis strategies lead to significantly different performance and that to achieve both high sensitivity and specificity in the detection of copy number using genome-wide array platforms, multiple parallel strategies must be considered. In terms of sensitivity, the best overall performances were obtained by HMM and CBS with $\alpha = 0.05$ with average FNR $= 0.252 \pm 0.101$ and 0.261 ± 0.114, respectively. When looking at subcategories of CNVs based on size and amplitude, we observed that the best FNRs are obtained for homozygous deletions and high amplitude gains, despite the CNV size, and that poorer FNRs are obtained for short CNVs (1–4 kb) and for small amplitudes, with CBS with $\alpha = 0.01$ and the 'undo' option scoring the highest FNR. When assessing specificity, the best overall performances were scored by GLAD ($d = 6$, 12) and CBS ($\alpha = 0.010$ and 'undo' option). Importantly, GLAD with $d = 6$ detects a higher number of segments (on average ~1.6 times the segments detected by GLAD with $d = 12$ or by CBS), therefore providing greater sensitivity with no apparent cost to specificity (Tables 1–4).

Comparing the Affymetrix and Agilent platforms, we investigated to what extent FNR and FPR performances differ for CNVs overlapping segmental duplication loci versus CNVs which do not. Importantly, segmental duplication regions are enriched for deletion and duplication events, specifically NAHR. When comparing performance obtained for segmental duplication regions relative to non-segmental duplication regions, we consistently detected overall increases in FNR and FPR across each algorithm (Wilcoxon test P-values ranging from 8×10^{-14} to 4×10^{-3}). This quantitative comparison supports the fact that segmental duplication regions are intrinsically harder to evaluate for CNV detection and assessment, because the relative signals for copy number differences are lower (Figure 3) and that Agilent signal-to-noise ratio tends to facilitate signal interpretation in complex regions.

It is worth highlighting that this study assessed CNV detection performances in a discovery-like fashion. When dealing with known CNV loci—known coordinates from previous studies—the assessment of copy number state in single samples is a completely different task. If highly specific probes are available, the assessment of single sample copy number state at known loci can be performed with as little as one marker. Supplementary Figure S7 shows the distribution of $\log_2$ ratios across HapMap samples included in this study by using single Affymetrix and Agilent markers (positions chr4:64 386 072 and chr4:64 381 900, respectively) located on a biallelic CNV (Variation_38050).

In conclusion, this study quantitatively measured the effect of experimental and analytical choices on CNV detection and characterization, providing insights into the highlights and limitations of using short sequence oligonucleotide arrays. Importantly, short sequence oligonucleotide arrays allow for the assessment of SNP genotypes, which can contribute to the overall study, such as population stratification (ethnicity) (22), inter-sample similarity and identity test (23) and inbreeding coefficient assessment (24); SNP probe single allele data can be utilized to validate copy number calls (see Supplementary Figure S8) (18), and SNP and CNV probe intensities can be jointly investigated in phenotype association studies. Although this study focused on data generated by two platforms (i.e. Affymetrix and Agilent), we envision that the FNR and FPR result trends we obtained are generalizable to other platforms and all evaluations related to the reference model effect and the segmental duplication regions are conceptually valid and platform independent. Whereas proper experimental design and CNV validation by PCR, qPCR, FISH or sequencing are required for a successful study, a lot can be done on the analysis side to exploit genome-wide CNV data and to enhance the sensitivity and specificity of the results based upon the goals of an individual study.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the MAQC-II CNV Analysis Team, Mark A. Rubin and Mark B. Gerstein for fruitful discussions. The authors would also like to thank Anya Tsalenko for detailed technical discussions and review of the manuscript, and Agilent Technologies for sharing unpublished data.

FUNDING

The Starr Foundation Cancer Consortium (to F.D. and S.B.); Clinical and Translation Science Center at Weill Cornell Medical College (Grant number UL1RR024996) (to F.D.). Funding for open access charge: Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York.

Conflict of interest statement. None declared.

REFERENCES

1. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat Genet*., **36**, 949–951.
2. Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M., Chi,M. *et al.* (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
3. de Smith,A.J., Tsalenko,A., Sampas,N., Scheffer,A., Yamada,N.A., Tsang,P., Ben-Dor,A., Yakhini,Z., Ellis,R.J., Bruhn,L. *et al.* (2007) Array CGH analysis of copy number

variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.*, **16**, 2783–2794.
4. McCarroll,S.A., Kuruvilla,F.G., Korn,J.M., Cawley,S., Nemesh,J., Wysoker,A., Shapero,M.H., de Bakker,P.I., Maller,J.B., Kirby,A. *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.
5. Shaikh,T.H., Gai,X., Perin,J.C., Glessner,J.T., Xie,H., Murphy,K., O'Hara,R., Casalunovo,T., Conlin,L.K., D'Arcy,M. *et al.* (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res.*, **19**, 1682–1690.
6. Gokcumen,O. and Lee,C. (2009) Copy number variants (CNVs) in primate species using array-based comparative genomic hybridization. *Methods*, **49**, 18–25.
7. Lee,A.S., Gutierrez-Arcelus,M., Perry,G.H., Vallender,E.J., Johnson,W.E., Miller,G.M., Korbel,J.O. and Lee,C. (2008) Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies. *Hum. Mol. Genet.*, **17**, 1127–1136.
8. Perry,G.H. (2008) The evolutionary significance of copy number variation in the human genome. *Cytogenet. Genome Res.*, **123**, 283–287.
9. Perry,G.H., Tchinda,J., McGrath,S.D., Zhang,J., Picker,S.R., Caceres,A.M., Iafrate,A.J., Tyler-Smith,C., Scherer,S.W., Eichler,E.E. *et al.* (2006) Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl Acad. Sci. USA*, **103**, 8006–8011.
10. McCarroll,S.A. and Altshuler,D.M. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.
11. Brennan,C., Zhang,Y., Leo,C., Feng,B., Cauwels,C., Aguirre,A.J., Kim,M., Protopopov,A. and Chin,L. (2004) High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res.*, **64**, 4744–4748.
12. Greshock,J., Feng,B., Nogueira,C., Ivanova,E., Perna,I., Nathanson,K., Protopopov,A., Weber,B.L. and Chin,L. (2007) A comparison of DNA copy number profiling platforms. *Cancer Res.*, **67**, 10173–10180.
13. Perry,G.H., Ben-Dor,A., Tsalenko,A., Sampas,N., Rodriguez-Revenga,L., Tran,C.W., Scheffer,A., Steinfeld,I., Tsang,P., Yamada,N.A. *et al.* (2008) The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.*, **82**, 685–695.
14. Hupe,P., Stransky,N., Thiery,J.P., Radvanyi,F. and Barillot,E. (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**, 3413–3422.
15. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
16. International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
17. Carter,N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39**, S16–S21.
18. Itsara,A., Cooper,G.M., Baker,C., Girirajan,S., Li,J., Absher,D., Krauss,R.M., Myers,R.M., Ridker,P.M., Chasman,D.I. *et al.* (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.*, **84**, 148–161.
19. Korn,J.M., Kuruvilla,F.G., McCarroll,S.A., Wysoker,A., Nemesh,J., Cawley,S., Hubbell,E., Veitch,J., Collins,P.J., Darvishi,K. *et al.* (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.
20. Demichelis,F., Setlur,S.R., Beroukhim,R., Perner,S., Korbel,J.O., Lafargue,C.J., Pflueger,D., Pina,C., Hofer,M.D., Sboner,A. *et al.* (2009) Distinct genomic aberrations associated with ERG rearranged prostate cancer. *Genes Chromosomes Cancer*, **48**, 366–380.
21. Weir,B.A., Woo,M.S., Getz,G., Perner,S., Ding,L., Beroukhim,R., Lin,W.M., Province,M.A., Kraja,A., Johnson,L.A. *et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature*, **450**, 893–898.
22. Price,A.L., Patterson,N.J., Plenge,R.M., Weinblatt,M.E., Shadick,N.A. and Reich,D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
23. Demichelis,F., Greulich,H., Macoska,J.A., Beroukhim,R., Sellers,W.R., Garraway,L. and Rubin,M.A. (2008) SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines. *Nucleic Acids Res.*, **36**, 2446–2456.
24. Leutenegger,A.L., Prum,B., Genin,E., Verny,C., Lemainque,A., Clerget-Darpoux,F. and Thompson,E.A. (2003) Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet.*, **73**, 516–523.